Investigating very deep highway networks for parametric speech synthesis

Investigating Very Deep Highway Networks for Parametric Speech Synthesis

Xin Wang[a,b,*], Shinji Takaki[a], Junichi Yamagishi[a,b,c]

[a]*National Institute of Informatics, 2-1-2, Hitotsubashi-cho, Chiyoda-ku, Tokyo, 101-8430, Japan.*
[b]*SOKENDAI, 2-1-2, Hitotsubashi, Chiyoda, Tokyo, 101-8430, Japan.*
[c]*The Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, EH8 9LW, United Kingdom.*

Abstract

Deep neural networks are powerful tools for classification and regression tasks. While a network with more than 100 hidden layers has been reported for image classification, how such a non-recurrent neural network with more than 10 hidden layers will perform for speech synthesis is as yet unknown. This work investigates the performance of deep networks on statistical parametric speech synthesis, particularly the question of whether different acoustic features can be better generated by a deeper network. To answer this question, this work examines a multi-stream highway network that separately generates spectral and F0 acoustic features based on the *highway* architecture. Experiments on the Blizzard Challenge 2011 corpus show that the accuracy of the generated spectral features consistently improves as the depth of the network increases from 2 to 40, but the F0 trajectory can be generated equally well by either a deep or a shallow network. Additional experiments on a single-stream highway and normal feedforward network, both of which generate spectral and F0 features from a single network, show that these networks must be deep enough to generate both kinds of acoustic features well. The difference in the performance of multi- and single-stream highway networks is further analyzed on the basis of the networks' activation and sensitivity to input features. In general, the highway network with more than 10 hidden layers, either multi- or single-stream, performs better on the experimental corpus than does a shallow network.

Keywords: Text-to-Speech, Statistical parametric speech synthesis, Deep neural network, Highway neural network

1. Introduction

Speech synthesis aims at creating natural-sounding speech waveforms and is used in various types of application with speech waveforms as output. A widely used application is Text-to-Speech (TTS) synthesis [1], where the speech is synthesized to read aloud the input text. Its social value is obvious in human-machine and human-human communication, e.g., when a disabled human speaker cannot articulate sounds.

TTS is difficult because of the ambiguous association between text and speech. Despite the recent trend towards end-to-end TTS [2], most existing TTS systems consist of individual front- and back-ends. The front-end infers the phonemic and prosodic information from the text, and then the back-end synthesizes the speech waveform given the output from the front-end. The TTS back-end, or the acoustic model, can be implemented on the basis of statistical parametric speech synthesis (SPSS), a framework that uses statistical models such as a hidden Markov model (HMM) to generate speech acoustic features and then construct the waveform [3, 4]. Recently, various acoustic models based on neural networks (NNs) have been proposed [5, 6, 7]. Some of these focus on the multi-mode distribution [8], cross-time dependence [9], or cross-dimension dependence [10] of compact and low-dimensional acoustic features. Other models focus on generating raw acoustic features such as the spectrum envelope [11, 12].

This work examines how the depth of a non-recurrent neural network affects its acoustic modeling performance, particularly on the commonly used low-dimensional spectral and F0 acoustic features. This is achieved by using a highway neural architecture. In our previous work, we proposed a multi-stream highway network, where multiple sub-networks were used to separately generate different acoustic features [13]. Experiments on a multi-stream highway network with 14 hidden layers showed that the sub-network for F0 was partially dormant, while the sub-network for spectral features was active. This indicated that the generation of F0 and spectral features may require networks with different depths. Motivated by the previous observation, this work specifically investigates whether different kinds of acoustic features can be better generated by a deeper network.

To answer the question, this work compares multi-stream highway networks of different depth when applied to the Blizzard Challenge 2011 corpus. The results show that spectral features can be better generated by a deeper network with up to 40 hidden layers, while F0 can be well generated by a network with 4 hidden layers. This work also includes experiments on single-stream highway and conventional feedforward networks, where all acoustic features are generated from a single network. It was found that the single-stream networks must be sufficiently deep to generate both spectral and F0 features well. These findings are further supported by a histogram analysis

*Corresponding author
Email addresses: wangxin@nii.ac.jp (Xin Wang), takaki@nii.ac.jp (Shinji Takaki), jyamagishi@nii.ac.jp (Junichi Yamagishi)

of each network's activation and the neurons' sensitivity to the input textual features. The results and analysis indicate that increasing the depth of the neural network is beneficial for the experimental NN-based acoustic models. Comparing the single- and multi-stream networks in a subjective test also shows that the performance of all the experimental networks is better when the depth is increased from 4 to 40.

This work is closely related to pioneering work using neural networks for acoustic modeling in TTS [6, 8], where networks with up to 7 hidden layers were reported. However, this work explores deeper networks. Not only the normal feedforward network, but also two types of highway network are considered. More importantly, this work provides both results and analysis regarding the single- and multi-stream networks. Although several recent TTS system also used complex and deep neural networks [2, 14, 15], these large systems work on high-dimensional acoustic features or waveform sampling points. We believe that highway networks examined in this paper and the results on conventional low-dimensional acoustic features can be useful for the common SPSS-based speech synthesizers.

Section 2 of this paper introduces the multi- and single-stream highway networks. Section 3 then introduces the tools to analyze the neural network. Section 4 explains the details of the experiments, including the performance of highway networks with different depths when modeling spectral and F0 features. The trained networks are further analyzed in section 5 based on the tools introduced in section 3. Section 6 discusses the remaining issues, and section 7 concludes.

2. Acoustic model based on highway networks

2.1. Neural-network-based acoustic model for TTS

This work focuses on the neural-network-based acoustic model for TTS systems with a pipeline structure [1]. This type of TTS system uses a front-end to derive the information on the pronunciation and prosody of the input text. Based on the linguistic features, the SPSS-based back-end generates the acoustic features and then constructs the waveform. For the common configuration in English TTS, the prosodic features may include the pitch accent of syllables and boundary tone of phrases, and the acoustic features may consist of F0 and compact spectral features of each speech frame [16]. The acoustic model can be implemented on the basis of feedforward neural networks, where the input linguistic features are transformed into acoustic features frame by frame [6].

The advantage of a deep neural network is its ability to extract structural features from data [17, 18, 19]. For example, a non-recurrent network with more than 100 hidden layers performed best in a recent image classification task [20]. For acoustic modeling, existing work has explored feedforward neural networks with up to 7 hidden layers [6, 8]. Although this showed that a deeper network improved the accuracy of generated acoustic features, the performance of a network with more than 10 hidden layers remains unknown. This work thus aims to investigate the performance of deeper non-recurrent networks by analyzing their performance and behavior.

2.2. Highway-network-based acoustic model

A deep neural network cannot be easily trained using the back-propagation algorithm and the random initialization strategy. Various methods have been proposed to facilitate the training of deep neural networks, including using better initialization methods based on unsupervised pre-training [21, 22] and unsaturated activation functions [23, 24]. Another thread of research focuses on the network architecture that alleviates the difficulty of network training. One new architecture is the *highway network* [25], which combats the gradient-vanishing problem. It has been shown that a deep highway network with more than 10 hidden layers can be well trained for speech recognition without requiring complicated engineering tunings [26]. Therefore, this work investigates deep highway networks for parametric speech synthesis because of their simplicity and effectiveness.

2.2.1. Highway block

The highway network consists of one or multiple highway blocks. In one highway block, the input linguistic feature vector $\boldsymbol{x}$ is transformed by a conventional feedforward layer as

$$\mathcal{H}(\boldsymbol{x}) = f(\boldsymbol{W}_H\boldsymbol{x} + \boldsymbol{b}_H). \tag{1}$$

Here, $f(\cdot)$ is the non-linear activation function, $\boldsymbol{b}_H$ is the bias vector and $\boldsymbol{W}_H$ is the transformation matrix. Furthermore, the highway block uses a *highway gate* to compute a control vector

$$\mathcal{T}(\boldsymbol{x}) = \sigma(\boldsymbol{W}_T\boldsymbol{x} + \boldsymbol{b}_T), \tag{2}$$

and then merges the transformed feature vector $\mathcal{H}(\boldsymbol{x})$ with the input $\boldsymbol{x}$ as the output of this highway block:

$$\boldsymbol{y} = \mathcal{T}(\boldsymbol{x}) \odot \mathcal{H}(\boldsymbol{x}) + [\boldsymbol{1} - \mathcal{T}(\boldsymbol{x})] \odot \boldsymbol{x}. \tag{3}$$

Here, $\odot$ denotes element-wise multiplication, and the sigmoid function $\sigma(x) = \frac{1}{1+e^{(-x)}}$ is used in the highway gate. The above computational flow in one highway block is plotted in Figure 1.

Parameters $\boldsymbol{W}_T$ and $\boldsymbol{b}_T$ in the highway gate are also trainable. When the output of the gate $\mathcal{T}(\boldsymbol{x})$ is approximately zero, the input $\boldsymbol{x}$ can be directly propagated forwards; i.e., $\boldsymbol{y} \approx \boldsymbol{x}$. In this case, the gradient can also be propagated backwards without being attenuated by the feedforward transformation layer in the highway block. Thus, a very deep network based on highway blocks can be trained by using the standard gradient-descent back-propagation algorithm. Note that $\mathcal{H}(\boldsymbol{x})$ can be a transformation conducted by multiple feedforward layers. In other words, one highway block can contain more than one feedforward transformation layer.

2.2.2. Single-stream and multi-stream highway networks

A highway network based on one or more highway blocks can be directly used as the acoustic model for TTS, which is shown on the right side of Figure 1. Because all the acoustic features are generated by a single network, we call it a *single-stream* highway network. The word 'stream' is borrowed from the HMM-based parametric framework [27]. Besides the single-stream architecture, we propose a multi-stream highway

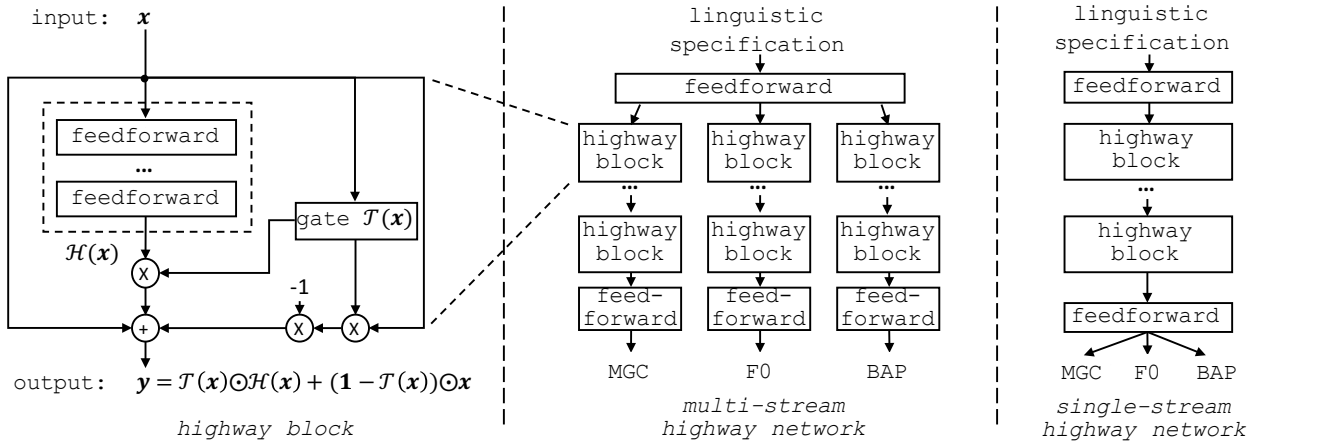


Figure 1: Computation flow in one highway block (left), multi- (middle) and single-stream (right) highway networks. MGC and BAP denote mel-generalized cepstral coefficients and band aperodicity. The 'feedforward' layer is a conventional non-linear or linear transformation layer in the feedforward neural network.

network as shown in the middle of Figure 1. On the input side, a linear projection layer transforms the input vector into a shared hidden vector. The multi-stream highway network then uses several sub-networks to separately transform the shared vector into different acoustic features.

There are two reasons for us to investigate multi-stream networks. First, the hidden feature vector in a single-stream network encodes the information for generating both spectral and F0 features, so the effect of the network's depth on the spectral and F0 features cannot be separately examined. In addition, the hidden vector may be biased toward high-dimensional spectral features and affect the performance of F0 generation. Thus, a multi-stream architecture is investigated to clearly show the result for each acoustic feature type.

Second, we want to compare the change of performance when the depth of single- and multi-stream networks is increased. Despite the claim that a single-stream network can model the correlation between spectral and F0 features [28], other research shows that the correlation between F0 and spectral features is somewhat weak, at least for reading speech in a neutral style [29, 30]. Recent work also shows that the F0 generation performance becomes worse even when spectral features are better generated on the basis of multi-task learning [31]. In addition, spectral and F0 features may rely on different input textual features [32], thus on different hidden representations in the network. Considering the above argument, we wonder that multi- and single-stream highway networks may perform differently on F0 and spectral feature generation as the depth is increased.

3. Tools to analyze the highway network

3.1. Histogram of the output of the highway gate

Each network's performance can be evaluated by calculating the accuracy of the generated acoustic features. In addition, we introduce two tools to analyze the highway network. The first tool is the histogram of the output of the highway gate; i.e., $\mathcal{T}(\boldsymbol{x})$ in Equation 2. As discussed in section 2.2.1, a highway block tends to avoid using the transformed $\mathcal{H}(\boldsymbol{x})$ as its output when the highway gate is almost closed ($\mathcal{T}(\boldsymbol{x}) \approx 0$). Thus, the histogram on $\mathcal{T}(\boldsymbol{x})$ can reflect the behavior of the highway block and thus the whole network. In practice, we activate the network with the input data for one phoneme and collect $\mathcal{T}(\boldsymbol{x})$ to plot the histogram. Note that the order of the feature dimension is ignored.

3.2. Sensitivity of a neuron to input textual features

The second tool evaluates the sensitivity of a neuron to the input textual feature. Suppose the input feature vector at time t takes the value s for a feature class $\boldsymbol{S}$. For example, when $\boldsymbol{S} = \{/a/, /t/, \cdots\}$ refers to the phoneme identity, $s = /a/$ means this frame realizes the phoneme $/a/$. Then, all the input vectors that take the same feature value s can be fed to the neural network, and the output of the k-th neuron can be summarized as

$$a_k(s) = \frac{exp(\tilde{a}(s))}{\sum_{s \in \boldsymbol{S}} exp(\tilde{a}(s))}, \tag{4}$$

where

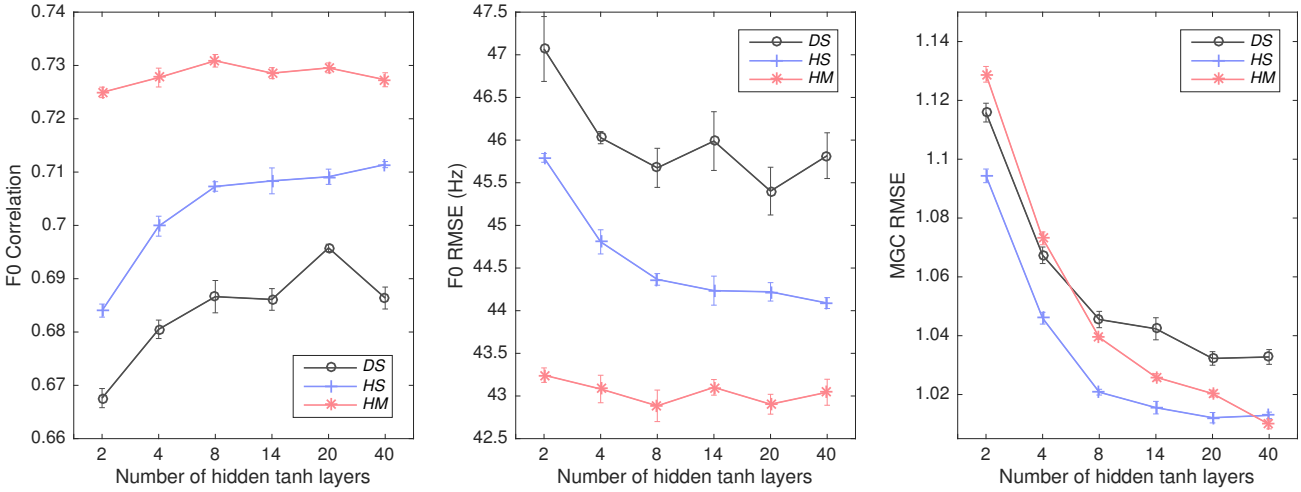
$$\tilde{a}_k(s) = \frac{\sum_t \gamma_s(t) h_k(t)}{\sum_t \gamma_s(t)}. \tag{5}$$

Here, $\gamma_s(t)$ is an indicator whose value is 1 if the input feature at time t takes the feature value s. On the basis of $a_k(s)$, the k-th neuron's sensitivity to the feature class $\boldsymbol{S}$ is defined as a normalized entropy over all possible feature values as

$$E_{\boldsymbol{S},k} = \frac{-\sum_{s \in \boldsymbol{S}} a_k(s) \log a_k(s)}{-\sum_{s \in \boldsymbol{S}} \frac{1}{|\boldsymbol{S}|} \log \frac{1}{|\boldsymbol{S}|}}, \tag{6}$$

where $|\boldsymbol{S}|$ is the number of possible values in class $\boldsymbol{S}$. Note that if $\boldsymbol{S}$ refers to a continuous feature such as syllable position, s is an index of a quantized interval. Intuitively, if $h_k(t)$ is constant $\forall s \in \boldsymbol{S}$, then $E_{\boldsymbol{S},k} = 1$ and the neuron is insensitive to the feature class S. The average sensitivity of a neuron group K can be further defined as $\bar{E}_{\boldsymbol{S}} = \frac{1}{|K|} \sum_{k \in K} E_{\boldsymbol{S},k}$.

This sensitivity measure was originally defined by Dr. Sim for a neural network in a speech recognition system [33]. The difference is that $\tilde{a}(s)$ in Equation 4 is divided by the number of matched data frames. Because the distribution of the textual features for TTS can be highly unbalanced, this division makes it possible to compare the entropy of different feature classes.



(a) F0 Correlation (b) F0 RMSE (Hz) (c) MGC RMSE

Figure 2: Performance of feedforward network (*DS*), single-stream (*HS*), and multi-stream highway network (*HM*) on test set when the depth of the network varies. Each highway block in *HS* and *HM* contains two hidden transformation layers with a *tanh* activation function. Each point is the average result over the last five training epochs of each network for two training trials. The standard deviation is shown as the error bar.

Table 1: Experimental networks.

	Definition
DS	Single-stream deep feedforward network
HS	Single-stream highway network
HM	Multi-stream highway network

4. Experiments

4.1. Corpus and network notation

Experiments used the Blizzard Challenge 2011 Nancy corpus that has 12072 English utterances [34]. Both the test and validation set contained 500 randomly selected utterances. Mel-generalized cepstral coefficients (MGCs) of order 60, continuous F0 trajectory, voiced/unvoiced (V/U) condition, and band aperiodicity (BAP) of order 25 were extracted for each speech frame by using the STRAIGHT vocoder [35]. The Flite toolkit [36] conducted the text-analysis for the entire corpus. The output of Flite were converted into a vector of order 382 as the input $\boldsymbol{x}_t$ to the neural network. This vector encodes common textual features similar to those used in the HMM-based framework [16]. Experiments were conducted on the three types of neural network listed in Table 1. The toolkit for training the neural network was modified on the basis of the CURRENNT library [37] [1].

4.2. Experiments on the depth of the neural network

4.2.1. Network configuration and training recipe

This experiment investigated how the depth of a multi-stream highway network influences its ability to generate different acoustic features. For reference, the single-stream highway and feedforward network were also examined.

All the highway blocks in the single-stream highway network (*HS*) had a layer size of 382, equal to the dimension of the input vector. Each highway block contained 2 hidden layers based on a *tanh* activation function. The bias of the highway gate was initialized as -1.5 according to the results of a preliminary experiment comparing -1.5, 0.0 and 1.5. The other parameters were initialized using the normalized initialization strategy [38] [2]. The single-stream feedforward network *DS* had a layer size of 382 and a *tanh* function for all layers. Its model parameters were initialized using the normalized initialization strategy.

The multi-stream highway network *HM* contained three sub-networks for MGC, F0, and BAP. The layer size of each sub-network was 256. The layer size of the linear projection layer near the input side was 768 ($= 256 * 3$). Each sub-network was configured and initialized in the same way as the *HS* network.

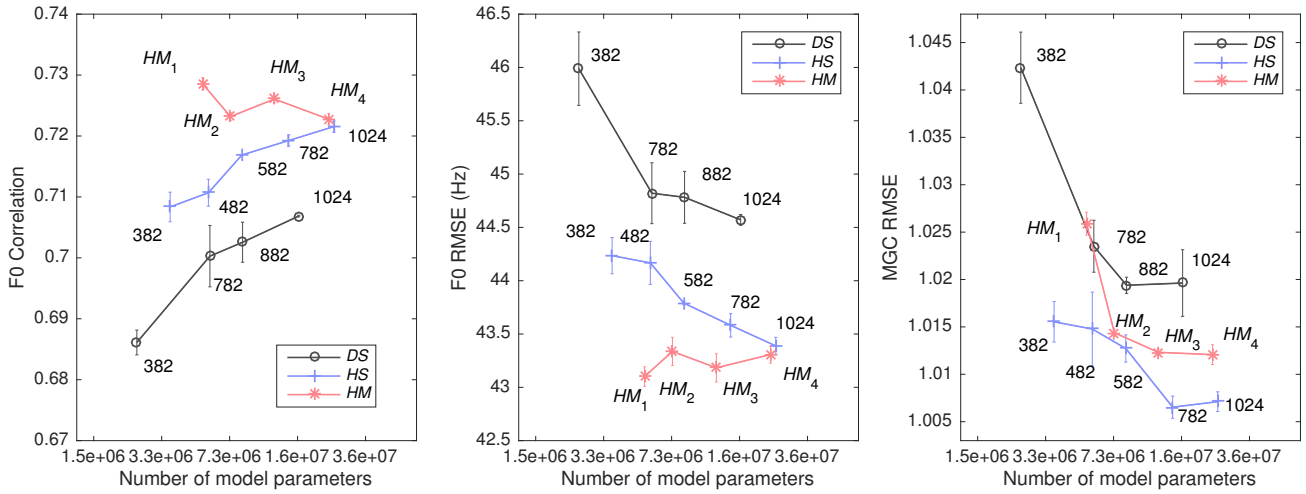
These three networks were trained with the number of *tanh*-based hidden layers set to $\{2, 4, 8, 14, 20, 40\}$. The the stochastic gradient descent method with early stopping was used for trainining. Batch normalization was not used [39] as our initial experiment showed that its effect is limited when the normalized initialization is used. The natural alignment on the test data was used in the objective evaluation. The objective measure was calculated on the test data for the last five training epochs of two training trails. The mean and standard deviation of each objective measure were then computed.

4.2.2. Results and analysis

The results are shown in Figure 2. Our first observation is that the MGC RMSE decreased when the depth of *HM* was increased from 2 to 40. The MGC RMSE curves of the other two experimental networks also decreased in a similar way. These results suggest that both single- and multi-stream networks are better at generating spectral features if they are deeper.

[1] The code and samples can be found at `http://tonywangx.github.io`.

[2] Our previous work [13] used Gaussian noise to initialize the highway network. However, we found that the normalized initialization strategy [38] led to a consistently better performing highway and feedforward networks.



(a) F0 Correlation (b) F0 RMSE(Hz) (c) MGC RMSE

Figure 3: Performance of multi-stream highway network (*HM*), the single-stream highway (*HS*) and feedforward network (*DS*) on the test set. Each network contains 14 hidden *tanh* layers. The number associated with a *HS* or *DS* network denotes the layer size. Table 2 defines HM_1 to HM_4.

However, the error curves on F0 showed different patterns. While the F0 curves of *HS* and *DS* start from a low point and gradually rise with increased network depth, the F0 curve of *HM* is almost flat. Particularly, *HM* with just 2 hidden *tanh* layers performed as well as the *HM* with 40 hidden layers in terms of F0. These results suggest that, for a network using *HM* architecture, more hidden layers is better for generating spectral features but not necessarily for F0.

Note that *HM* had a larger total layer size than *HS* and *DS*, which may influence *HM*'s performance on F0 especially when the network was shallow. This possibility is examined in the next section. However, the shape of the RMSE curves generally shows how the F0 and spectral feature generation performs differently in the multi-stream network.

4.3. Experiments on the layer size of the neural network

4.3.1. Network configuration and training recipe

This experiment investigated network performance given a fixed depth but different layer sizes. All the neural networks in Table 1 were trained with 14 hidden layers. The depth of 14 was selected because a deeper network with a large layer size required too much GPU memory resource because of the current software implementation. For *HS* and *DS*, the trained networks in Section 4.2 with the layer size 382 were included for comparison. Meanwhile, *HS*s with layer sizes of 482, 582, 782 and 1024 were trained. Then, *DS*s with layer sizes of 782, 882 and 1024 were trained so that they were comparable with the *HS*s in terms of model size. The *HM*s used the configurations shown in Table 2. The training recipe was the same as in Section 4.2. The total number of model parameters of each network was collected in order to compare results across different network types.

4.3.2. Results and analysis

According to the results shown in Figure 3, *HM*'s F0 performance did not change much when the layer size of the F0 sub-network changed. In contrast, *HS* and *DS* performed better on F0 as their layer size grew larger. On MGC, all three types of network performed better with a larger layer size.

Table 2: Network structure of *HM* networks in Figure 3.

	Layer size of the sub-network		
	MGC stream	F0 stream	BAP stream
HM_1	256	256	256
HM_2	382	256	256
HM_3	512	382	256
HM_4	768	512	256

These results are consistent with those in Section 4.2. For the multi-stream network, a larger network is better at spectral feature modeling than a smaller network, but it might not surpass a shallower network on F0 modeling. A larger single-stream network performs better on both acoustic features. Particularly, its F0 generation performance is similar to that of the multi-stream network. A single-stream network might be dominated by the hidden features for the spectral features. If that is the case, only in a very large network would some of the nodes in the single-stream network be assigned to model the F0. This possibility will be further discussed in Section 5.1.

4.4. Subjective evaluation

The objective evaluation results showed that spectral features can be better modeled by a very deep network while this is not the case for F0. However, it is of interest to know whether the difference is perceptible. Thus, before analyzing the network in detail, we describe a subjective evaluation in this section. We prepared 5 groups of synthetic samples in Table 3 and organized a MUSHRA test [40] in CSTR of the University of Edinburgh. Ten paid native English speakers participated in the test. Among the experimental groups, a comparison of *H*1, *H*2, and *H*3 was used to evaluate the perceptible difference in the generated spectral features, while a comparison of *H*1, *H*4, and *H*5 was used to show the difference in the generated F0s.

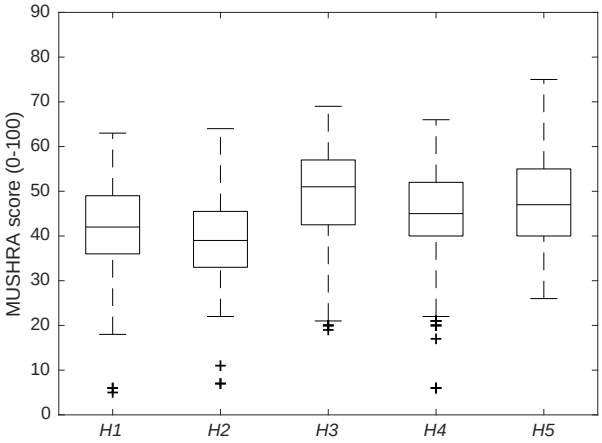


Figure 4: Results of the MUSHRA test on synthetic speech based on acoustic features generated by several multi-stream highway networks (HM_1). The central mark is the median, and the edges of the box indicate the 25th and 75th percentiles. Each group is defined in Table 3.

Table 3: Summary of test groups in Figure 4. HM_1 used 256 as the layer size for each sub-network (Table 2). The number after HM_1 denotes the number of hidden tanh layers.

	Network for generating the acoustic feature	
	MGC & BAP	F0
*H*1	HM_1 14 hidden layers	HM_1 14 hidden layers
*H*2	HM_1 4 hidden layers	HM_1 14 hidden layers
*H*3	HM_1 40 hidden layers	HM_1 14 hidden layers
*H*4	HM_1 14 hidden layers	HM_1 4 hidden layers
*H*5	HM_1 14 hidden layers	HM_1 40 hidden layers

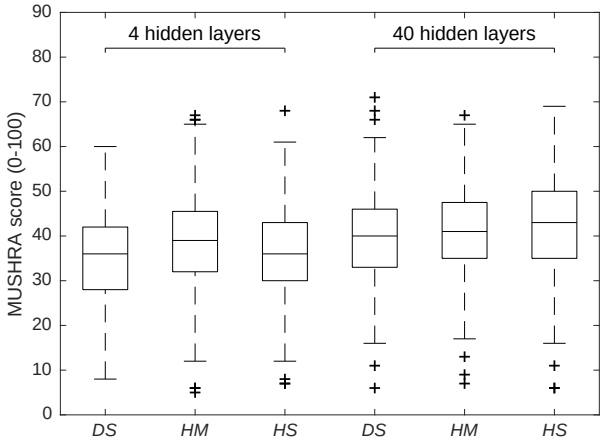


Figure 5: Results of the MUSHRA test on the single-stream feedforward network (*DS*), single-stream highway network (*HS*), and multi-stream highway network (HM_1). The central mark is the median, and the edges of the box are the 25th and 75th percentiles. HM_1 used 256 as the layer size for each sub-network (Table 2).

Table 4: Results of the two-sided Wilcoxon signed rank tests ($\alpha = 0.01$) between groups in Figure 5. The number denotes the depth of the network. ■ indicates a significant difference, while □ indicates an insignificant difference.

	DS 4	*HM* 4	*HS* 4	*DS* 40	*HM* 40	*HS* 40
DS 4	-	■	□	■	■	■
HM 4	■	-	■	□	□	■
HS 4	□	■	-	■	■	■
DS 40	■	□	■	-	□	■
HM 40	■	□	■	□	-	□
HS 40	■	■	■	■	□	-

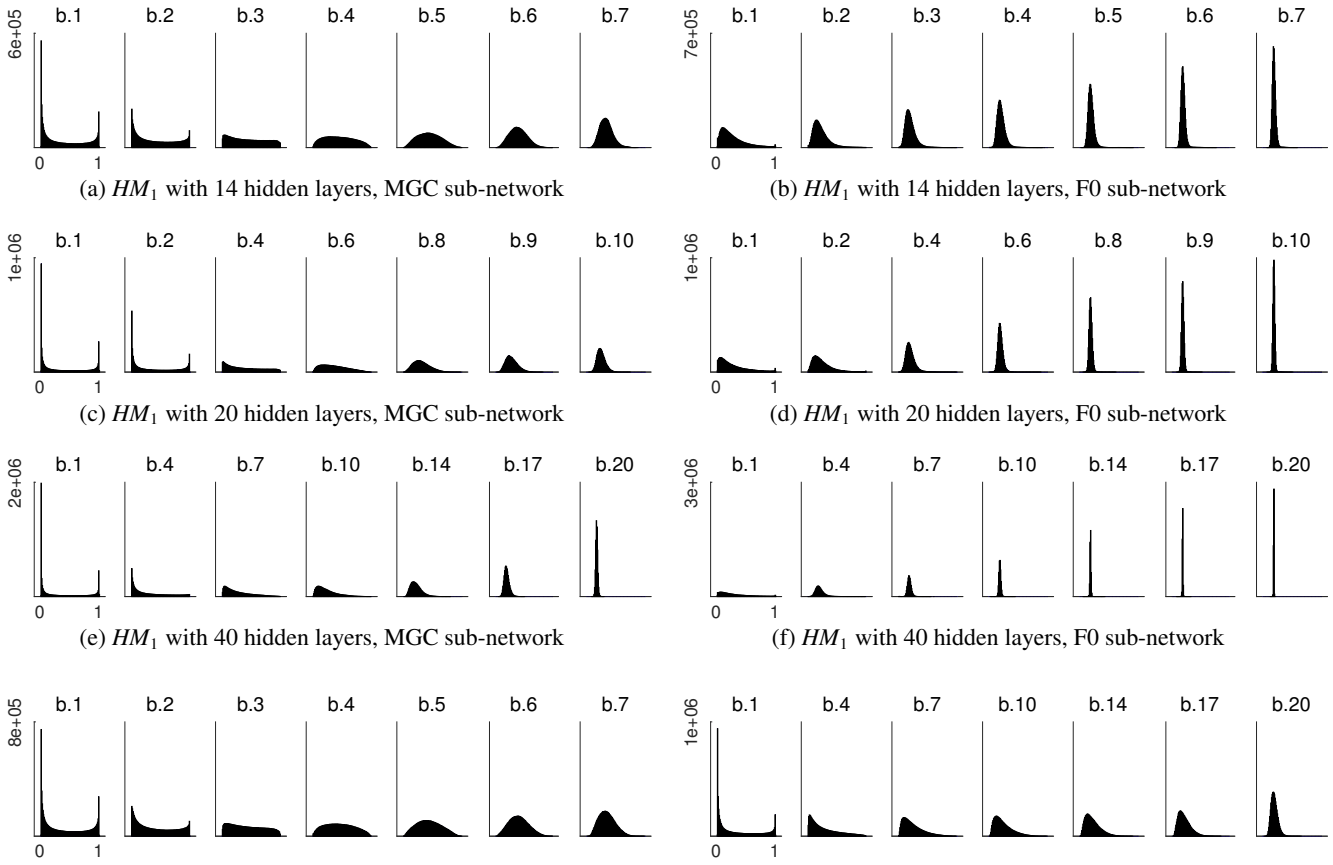
The results are shown in Figure 4. Regarding the spectral features, *H*3 achieved the best performance, followed by *H*1 and then *H*2. A two-sided *t*-test showed that the average score of *H*3 was statistically higher than that of *H*1 ($p \approx 0.000$) and *H*2 ($p \approx 0.000$). Meanwhile, *H*1's average score was higher than that of *H*2 ($p = 0.049$). These results indicated that a deeper HM_1 network generated spectral features that were perceived to be better in quality.

Regarding F0 among the three groups with different configurations, *H*1 had a lower score than *H*4 ($p = 0.003$), even though *H*1 used F0 generated by a network with 10 more hidden layers. Similarly, *H*5 was not significantly different from *H*4 ($p = 0.088$) even though *H*5 used F0 from the deepest network with 40 hidden layers. Interestingly, although the objective evaluation of F0 in Figure 2 did not show huge differences among networks with different depths, the subjective evaluation showed a perceptible difference between HM_1 with 14 hidden layers and the other cases. This could be because different models generate quite different F0 trajectories even though the average RMSE and CORR metrics on the generated F0 trajectories are not so different. However, this result at least shows that a deeper network based on HM_1 does not guarantee improvement in the generated F0 trajectories, unlike the case for spectral features.

On the basis of the first MUSHRA test, the second test examined the overall performance of *DS*, *HS* and HM_1 with depths 4 and 40. The baseline depth was selected as 4 because it was commonly used for baseline feedforward neural networks. Here, the F0 and spectral features of one synthetic sample were generated by the same network.

The results of the second test are shown in Figure 5. Additionally, a two-sided Wilcoxon signed rank test is used to measure the significance of the difference and the results are shown in Table 4. For all types of experimental network, the synthetic samples generated by the network with 40 hidden layers had a higher average score than those of the same type of network with just 4 hidden layers. According to Table 4, increasing the depth to 40 led to a statistically significant improvement for *DS* and *HS*. This result is consistent with the objective evaluation and suggests that using a very deep network with more than 10 hidden layers is better than using a network with 4 hidden layers. Even the classical feedforward network is improved if it is sufficiently deep. When the network is very deep, both single- and multi-stream highway networks are good choices.

Interestingly, *HM* had a higher average score than *HS* when they each had 4 hidden layers. However, *HS* outperformed *HM* when the number of hidden layers was 40 for each network. One reason is that, when the network is shallow, the multi-stream structure takes advantage of the larger layer size. However, when the network is deep enough, even a single-stream network with a smaller layer size has enough capacity to model the spectral features as well as the F0 features.



(a) HM_1 with 14 hidden layers, MGC sub-network (b) HM_1 with 14 hidden layers, F0 sub-network

(c) HM_1 with 20 hidden layers, MGC sub-network (d) HM_1 with 20 hidden layers, F0 sub-network

(e) HM_1 with 40 hidden layers, MGC sub-network (f) HM_1 with 40 hidden layers, F0 sub-network

(g) *HS* with 14 hidden layers, 382 nodes per layer (h) *HS* with 40 hidden layers, 382 nodes per layer

Figure 6: Histogram of highway gates' output $\mathcal{T}(\boldsymbol{x})$ given input test data for phoneme /a/. Each highway block had 2 hidden *tanh* layers. Only 7 blocks are shown here for the network with more than 14 hidden layers. **b.1** always denotes the first block near the highway network's input side.

5. Analysis of the highway network

5.1. Activity of the highway network

The experiments in Section 4 showed that the depth of a multi-stream network had different effects on the generation of F0 and spectral features. To explain this result, we used the histograms introduced in Section 3 to analyze the network. Specifically, we used the input feature vectors of a specific phoneme from the test data as an excitation and plotted the histogram over the output of the highway gate. The histograms for different phonemes were similar and only the results for phoneme /a/ are shown here.

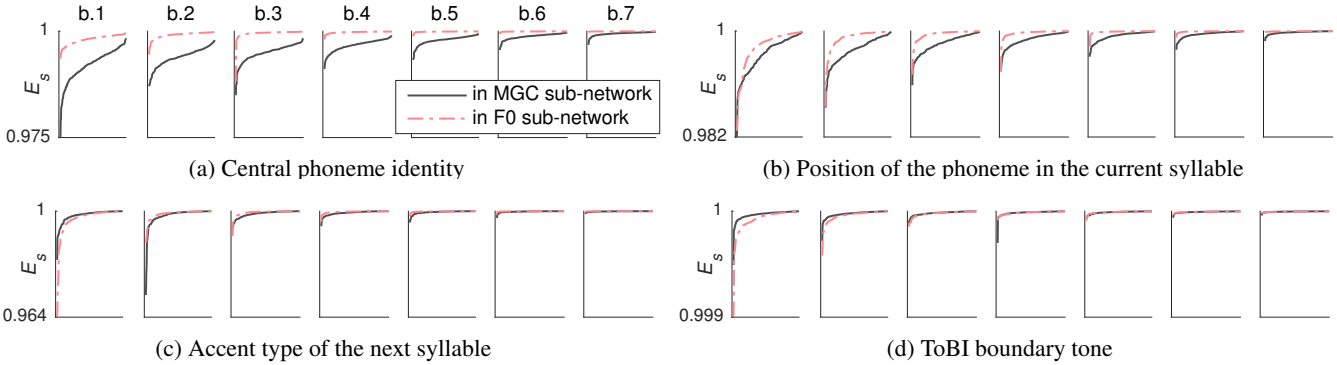
Figures 6 (a) and (b) show the histograms for HM_1 with 14 hidden layers. The histogram of the first block in the MGC sub-network, which is plotted in figure (b.1), showed an unbalanced binomial distribution, meaning that most of the gates were closed ($\mathcal{T}(\boldsymbol{x}) \approx 0$) while some gates were open ($\mathcal{T}(\boldsymbol{x}) \approx 1$). In the second block, the binomial distribution was kept to some degree. For blocks near the output side, the shape of the histogram gradually changed into a bell shape, which indicates that these blocks conducted complex transformations by summing the non-linear transformation output and the input.

The histograms for the F0 sub-network showed different patterns from those for the MGC sub-network. The histograms for the F0 sub-network, especially those near the end of the network, were spike-shaped. The width of the spike - variance of the data - was much smaller than that for MGC. Note that the spike in the histogram was located around 0.2, and this location was determined by the initial value of the gate units ($\frac{1}{1+exp(1.5)}$). Nevertheless, these blocks generally avoided using a non-linear transformation and simply delivered their input to the next block. Therefore, most of the blocks in the F0 sub-network were inactive.

Similar results were also observed in the histograms for HM_1 with 20 and 40 hidden layers, which are shown in Figure 6 (c-f). Particularly, the blocks near the output side in the deeper F0 sub-network had a sharper spike in the histogram. This result was consistent for all phonemes [3]. Because an inactive highway block tends to directly deliver the input data to the output side, adding these inactive blocks may not introduce additional feature transformation. This may explain why adding highway blocks did not improve the overall performance for F0 generation in the multi-stream highway network.

In contrast, the highway blocks in the MGC sub-network are generally active even in the network with 20 highway blocks. This result may explain the better performance of HM_1 in MGC generation as the depth increased from 2 to 40. Interestingly, Figure 6(e) also shows that the last block (b.20) in the MGC

[3]The other histograms can be found on `http://tonywangx.github.io`.



(a) Central phoneme identity

(b) Position of the phoneme in the current syllable

(c) Accent type of the next syllable

(d) ToBI boundary tone

Figure 7: The sensitivity measure $E_{S,k}$ for highway gate neurons in the MGC (solid line) and F0 (dash line) sub-networks of HM_1 with 14 hidden layers. **b.1** always denotes the first block near the network's input side. The horizontal axis, which ranges from 1 to 256, is the sorted index of the gate neurons according to $E_{S,k}$. A higher value of $E_{S,k}$ denotes less sensitivity.

Table 5: The most- and least-sensitive input features in the HM_1 network with 14 hidden layers. A lower $\bar{E}_S$ (average over neurons) reflects a more sensitive feature.

	in the MGC sub-network		*in the F0 sub-network*	
	feature class	$\bar{E}_S$	feature class	$\bar{E}_S$
Most sensitive	Phoneme identity	0.9923	Position of phoneme in syllable	0.9973
	Position of phoneme in syllable	0.9961	Accent type of next syllable	0.9974
	Position of phoneme in syllable (backward)	0.9967	Accent type of previous syllable	0.9980
	Number of previous stressed syllables in phrase	0.9974	Position of syllable in the word	0.9980
	Number of stressed syllables remained in phrase	0.9975	Phoneme identity	0.9981
Least sensitive	Position of phrase in utterance	0.9999	Number of words in previous phrase	0.9999
	Number of words in phrase	0.9999	Number of words in next phrase	0.9999
	Number of phrases in utterance	0.9999	ToBI boundary tone	0.9999
	Number of syllables in previous phrase	1.0000	Number of syllables in next phrase	0.9999
	ToBI boundary tone	1.0000	Number of syllables in previous phrase	1.0000

sub-network has a sharp histogram. This suggests that HM_1 might not require 20 highway blocks on the corpus.

The above analysis concerns the multi-stream highway network. Next, we plotted the histograms for *HS* with 14 hidden layers in Figure 6(g). In this case, we cannot differentiate the histograms for MGC and F0. Interestingly, the histograms of *HS* resemble those in the MGC sub-network of HM_1. This resemblance suggests that *HS* mainly focused on MGC modeling instead of F0. If this is true, it explains why *HS*'s performance on F0 was worse than that of *HM*.

The objective results in Section 3 showed that *HS* performed better on F0 when the network was deeper. One reason for this may be that a larger network has additional capacity for modeling F0. This assumption is supported to some extent by the histogram of the deepest *HS* network shown in Figure 6(h). Compared with b.20 in Figure 6(e), b.20 of *HS* didn't produce a sharp histogram. This indicates that the *HS* network took full advantage of the increased model capacity to model both F0 and spectral features.

5.2. *Sensitivity of the network to input textual features*

The previous section showed different levels of activity in the MGC and F0 sub-networks. Intuitively, we would expect a highway block to be active when the input feature correlates with the target feature and the mapping between them is complex. It is possible that the blocks for F0 were inactive because the target F0 had fewer dimensions than those of the spectral features. It is also possible that the input features were less correlated with F0.

Given the trained networks, we next examined the second possibility by analyzing the sensitivity of neurons to the input textual features. The HM_1 network trained with 14 hidden layers was analyzed based on the measure $E_{S,k}$ introduced in Section 3. This measure was calculated over the test set for every highway gate neuron and textual feature class. Additionally, the average $\bar{E}_S$ over the sub-network was calculated for each textual feature class S. The feature classes with the highest and lowest $\bar{E}_S$ value are listed in Table 5.

As Table 5 shows, the neurons in the MGC sub-network were the most sensitive to the phoneme identity and position of the phoneme in the syllable. The $E_{S,k}$ of each neuron with respect to these two feature classes are shown in Figure 7(a) and (b). The results are not suprising because the spectral features correlate mainly with the segmental information of the text. The results also showed that neurons in the F0 sub-network were sensitive to the segmental information. However, although F0 does correlate with segmental features [41, 42], we cannot see any input features above the word level that were highly ranked in Table 5. What's more, the ToBI boundary tone [43] was ranked as one of the least sensitive features. This result may be because the input features over the whole corpus, including the ToBI boundary tone, were automatically annotated by the text-analyzer. These input features may be too noisy to provide useful information for F0 generation.

A comparison of the curves for the MGC and F0 sub-

networks in Figure 7 generally indicates that the neurons in the F0 sub-network were less sensitive to all classes of input features. In addition, it shows that most of the neurons in the F0 sub-network were insensitive even to the top two feature classes. This may explain why most histograms of the highway blocks in the F0 sub-network had a bell shape. In contrast, the neurons in the MGC sub-network showed high sensitivity at least to the phoneme identity, including the neurons near the network's output. This is consistent with the shape of the histograms in the MGC sub-network.

In general, the results based on the sensitivity measure are consistent with the histograms of the MGC sub-network. This indicates that the network should be deep enough to fully transform the input features. However, the sufficient depth varies for different types of acoustic feature. Typically, the network for F0 can be shallower than the one for MGC.

6. Discussion

Experiments and analysis in this paper didn't measure the correlation between F0 and spectral features directly. However, the experiments' results at least suggest that F0 and spectral features cannot be assumed to be highly correlated for any corpus. They also suggest that the network structure should be carefully chosen to model spectral and F0 features together.

The experiments on the multi-stream highway network indicated that the F0 generation didn't gain from using a deeper network. Analysis in Section 5.2 suggested that one reason may be the uninformative input features. Particularly, the automatically inferred ToBI boundary tones turned out to be uninformative. Of course, this result depends on the language and the front-end, and it does not deny the effectiveness of ToBI boundary tone in other scenarios. For example, we analyzed a multi-stream highway network trained on a large Japanese corpus [44] and found that various information, such as the pitch accent type and inflected form of the word, were more useful than the phoneme identity since these features can be easily acquired from the surface text or retrieved from the lexicon. This result can be found online: http://tonywangx.github.io. However, even though the input features can be manually corrected, some of these features are not highly related to the realization of F0 [32]. As F0 is argued to rely on contextual information at various levels [45], it may be necessary to introduce more effective context features [46, 47]. Another approach is to consider the hierarchical property of the F0 and using hierarchical feature representation of F0 [48] and complex network structures [49].

Note that, although the depth of the experimental networks has been increased to 40, the computational resource required to train the network is still acceptable if the network is narrow. This was also supported by the work on highway networks for speech recognition [26]. From our experience, the total training time for the deepest multi-stream highway network was still short than that of training a neural network with just two feedforward, two recurrent layers and less model parameters.

Another interesting direction for further work is to introduce the highway connection to the recurrent neural networks (RNN). Recent research on this topic proposes highway RNNs with complex network structures [50, 51]. Based on a recent analysis on the highway network [52], we tried a simple approach to combine the highway and RNN. Although the experiment results are preliminary, they suggest that this structure performed relatively better than a deep RNN network with a similar number of model parameters. Furthermore, the training time was less than half of that for the deep RNN. To keep the focus of the idea and control the length of this paper, the work on the highway RNN is described in another report, which can be found on `http://tonywangx.github.io`.

7. Conclusion

This work aimed at showing the effect of a neural network's depth on the performance of F0 and spectral feature modeling. The comparison of multi-stream highway networks with different depths showed that the quality of generated spectral features improved when the network's depth was increased up to 40. However, the generated F0 from the network with 4 hidden layers was not significantly worse than that from a deeper network. Histogram-based analysis of the highway gate output in multi-stream highway networks showed that the highway blocks in the F0 sub-network behaved differently from those in the sub-network for spectral features. Typically, the highway blocks in the F0 sub-network tended to carry their input directly to the next layer, while blocks in the spectral sub-network conducted more complex non-linear transformations. This analysis supported the finding that the F0 sub-network was not required to be deep. For the commonly used single-stream architecture, the experiments showed that both the single-stream highway and conventional feedforward networks must be deep enough to provide sufficient network capacity to accurately model both F0 and spectral features. This finding was also supported by the histogram-based analysis.

In a nutshell, although a deeper network may be unnecessary for F0 modeling in the case of multi-stream highway networks, experiments and analysis on the experimental corpus suggest that a deeper network generally enables better parametric speech synthesis.

8. Acknowledgement

The work presented in this paper was partially supported by EPSRC EP/J002526/1 (CAF), by the Core Research for Evolutional Science and Technology (CREST) from the Japan Science and Technology Agency (JST) (uDialogue project), by MEXT KAKENHI Grant Numbers (26280066, 26540092, 15H01686, 15K12071, 16H06302, 16K16096) and by The Telecommunications Advancement Foundation Grant.

References

[1] P. Taylor, Text-to-Speech Synthesis, Cambridge University Press, 2009.
[2] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, et al., Deep voice: Real-time neural text-to-speech, arXiv preprint arXiv:1702.07825.

[3] H. Zen, K. Tokuda, A. W. Black, Statistical parametric speech synthesis, Speech Communication 51 (2009) 1039–1064.

[4] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, K. Oura, Speech synthesis based on hidden Markov models, Proceedings of the IEEE 101 (5) (2013) 1234–1252.

[5] Z. H. Ling, S. Y. Kang, H. Zen, A. Senior, M. Schuster, X. J. Qian, H. M. Meng, L. Deng, Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends, IEEE Signal Processing Magazine 32 (3) (2015) 35–52.

[6] H. Zen, A. Senior, M. Schuster, Statistical parametric speech synthesis using deep neural networks, in: Proc. ICASSP, 2013, pp. 7962–7966.

[7] Y. Fan, Y. Qian, F. Xie, F. K. Soong, TTS synthesis with bidirectional LSTM based recurrent neural networks, in: Proc. Interspeech, 2014, pp. 1964–1968.

[8] H. Zen, A. Senior, Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis, in: Proc. ICASSP, 2014, pp. 3844–3848.

[9] K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda, Trajectory training considering global variance for speech synthesis based on neural networks, in: Proc. ICASSP, 2016, pp. 5600–5604.

[10] B. Uria, I. Murray, S. Renals, C. Valentini-Botinhao, J. Bridle, Modelling acoustic feature dependencies with artificial neural networks: Trajectory-RNADE, in: Proc. ICASSP, 2015, pp. 4465–4469.

[11] Z.-H. Ling, L. Deng, D. Yu, Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis, IEEE Transactions on Audio, Speech, and Language Processing 21 (2013) 2129–2139.

[12] S. Takaki, J. Yamagishi, A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis, in: Proc. ICASSP, 2016, pp. 5535–5539.

[13] X. Wang, S. Takaki, J. Yamagishi, Investigating very deep highway networks for parametric speech synthesis, in: Proc. SSW9, 2016, pp. 181–186.

[14] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499.

[15] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, R. A. Saurous, Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model, ArXiv e-printsarXiv:1703.10135.

[16] K. Tokuda, H. Zen, A. W. Black, An HMM-based speech synthesis system applied to english, in: Proc. SSW, 2002, pp. 227–230.

[17] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (1) (2009) 1–127.
URL http://dx.doi.org/10.1561/2200000006

[18] R. Eldan, O. Shamir, The power of depth for feedforward neural networks, in: Proc. Annual Conference on Learning Theory, 2016, pp. 907–940.

[19] M. Telgarsky, Benefits of depth in neural networks, in: Proc. Annual Conference on Learning Theory, 2016, pp. 1517–1539.

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385.
URL http://arxiv.org/abs/1512.03385

[21] G. E. Hinton, R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[22] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, S. Bengio, Why does unsupervised pre-training help deep learning?, The Journal of Machine Learning Research 11 (2010) 625–660.

[23] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: Proc. AISTATS, Vol. 15, 2011, pp. 315–323.

[24] V. Nair, G. E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proc. ICML, 2010, pp. 807–814.

[25] R. K. Srivastava, K. Greff, J. Schmidhuber, Highway networks, in: Proc. Deep Learning Workshop, 2015.

[26] L. Liang, R. Steve, Small-footprint deep neural networks with highway connections for speech recognition, in: Proc. Interspeech, 2016, pp. 12–16.

[27] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, in: Proc. Eurospeech, 1999, pp. 2347–2350.

[28] B. Chen, Z. Chen, J. Xu, K. Yu, An investigation of context clustering for statistical speech synthesis with deep neural network., in: Proc. Interspeech, 2015, pp. 2212–2216.

[29] Z. H. Ling, W. Zhang, R. Wang, Cross-stream dependency modeling for HMM-based speech synthesis, in: Proc. ISCSLP, 2008, pp. 1–4.

[30] O. Watts, G. E. Henter, T. Merritt, Z. Wu, S. King, From HMMs to DNNs: where do the improvements come from?, in: Proc. ICASSP, 2016, pp. 5505–5509.

[31] Z. Wu, C. Valentini-Botinhao, O. Watts, S. King, Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis, in: Proc. ICASSP, 2015, pp. 4460–4464.

[32] O. Watts, J. Yamagishi, S. King, The role of higher-level linguistic features in HMM-based speech synthesis, in: Proc. Interspeech, 2010, pp. 841–844.

[33] K. C. Sim, On constructing and analysing an interpretable brain model for the dnn based on hidden activity patterns, in: Proc. ASRU, 2015, pp. 22–29.

[34] S. King, V. Karaiskos, The Blizzard Challenge 2011, in: Proc. Blizzard Challenge Workshop, 2011, pp. 1–10.

[35] H. Kawahara, I. Masuda-Katsuse, A. d. Cheveigne, Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds, Speech Communication 27 (1999) 187–207.

[36] HTS Working Group, The English TTS system Flite+HTS_engine (2014).
URL http://hts-engine.sourceforge.net/

[37] F. Weninger, J. Bergmann, B. Schuller, Introducing CURRENT: The Munich open-source CUDA recurrent neural network toolkit, The Journal of Machine Learning Research 16 (1) (2015) 547–551.

[38] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proc. AISTATS, 2010, pp. 249–256.

[39] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proc. ICML, 2015, pp. 448–456.

[40] Method for the subjective assessment of intermediate quality level of coding systems, in: ITU-R BS.1534-1, International Telecommunication Union Radiocommunication Assembly, 2003, http://www.itu.int/rec/R-REC-BS.1534.
URL http://www.itu.int/rec/R-REC-BS.1534

[41] K. E. A. Silverman, The Structure and Processing of Fundamental Frequency Contours, Ph.D. thesis, University of Cambridge (1987).

[42] D. H. Whalen, A. G. Levitt, The universality of intrinsic F0 of vowels, Journal of phonetics 23 (3) (1995) 349–366.

[43] K. E. A. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, J. Hirschberg, ToBI: a standard for labeling English prosody, in: Proc. ICSLP, 1992, pp. 867–870.

[44] H. Kawai, T. Toda, J. Yamagishi, T. Hirai, J. Ni, N. Nishizawa, M. Tsuzaki, K. Tokuda, Ximera: A concatenative speech synthesis system with large scale corpora, IEICE Transactions on Information and System (Japanese Edition) (2006) 2688–2698.

[45] J. Cole, Prosody in context: a review, Language, Cognition and Neuroscience 30 (1-2) (2015) 1–31.

[46] R. Dall, K. Hashimoto, K. Oura, Y. Nankaku, K. Tokuda, Redefining the linguistic context feature set for HMM and DNN TTS through position and parsing, in: Proc. Interspeech, 2016, pp. 2851–2855.

[47] H. Hashimoto, K. Hirose, N. Minematsu, Context labels based on "bunsetsu" for HMM-based speech synthesis of Japanese, in: Proc. SSW8, 2013, pp. 35–39.

[48] A. S. Suni, D. Aalto, T. Raitio, P. Alku, M. Vainio, et al., Wavelets for intonation modeling in HMM speech synthesis, in: Proc. SSW8, 2013, pp. 285–290.

[49] M. S. Ribeiro, O. Watts, J. Yamagishi, Syllable-level representations of suprasegmental features for DNN-based text-to-speech synthesis, Proc. Interspeech.

[50] J. G. Zilly, R. K. Srivastava, J. Koutník, J. Schmidhuber, Recurrent highway networks, arXiv preprint arXiv:1607.03474.

[51] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, J. Glass, Highway long short-term memory rnns for distant speech recognition, in: Proc. ICASSP, 2016, pp. 5755–5759.

[52] K. Greff, R. K. Srivastava, J. Schmidhuber, Highway and residual networks learn unrolled iterative estimation, arXiv preprint arXiv:1612.07771.